1. tocsv.py 选取有用属性，提取数据写入data\_c.csv文件，并且对不合理点、离群点进行过滤。其中部分属性attrs = ['decoration\_condition', 'elevator', 'ownership','framework']中存在unknown值，起初使用-1代替，作为新的类型，模型预测效果不够理想；后使用随机森林RandomForestClassifier对unknown值进行预测，处理后的数据写入data\_r.csv文件。

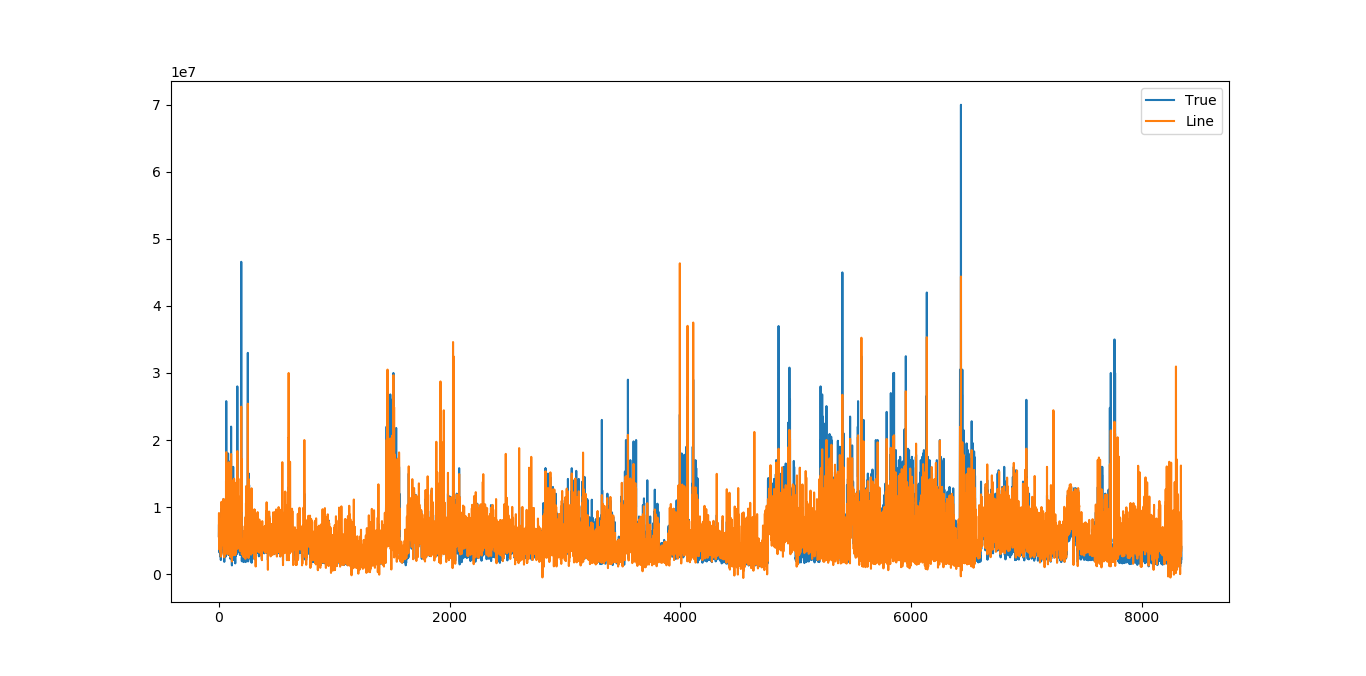
（27个输入属性 + 2个输出属性 平均房价和总房价，预测时预测总房价）

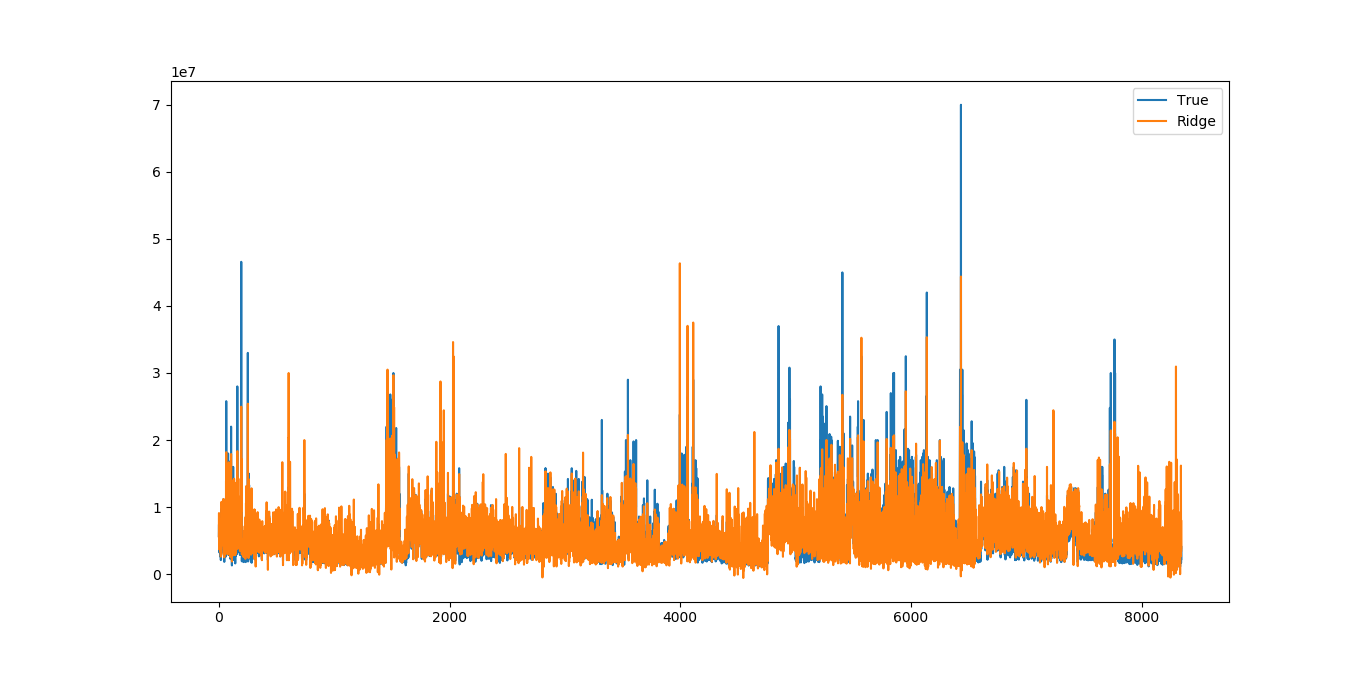
（61609条数据）

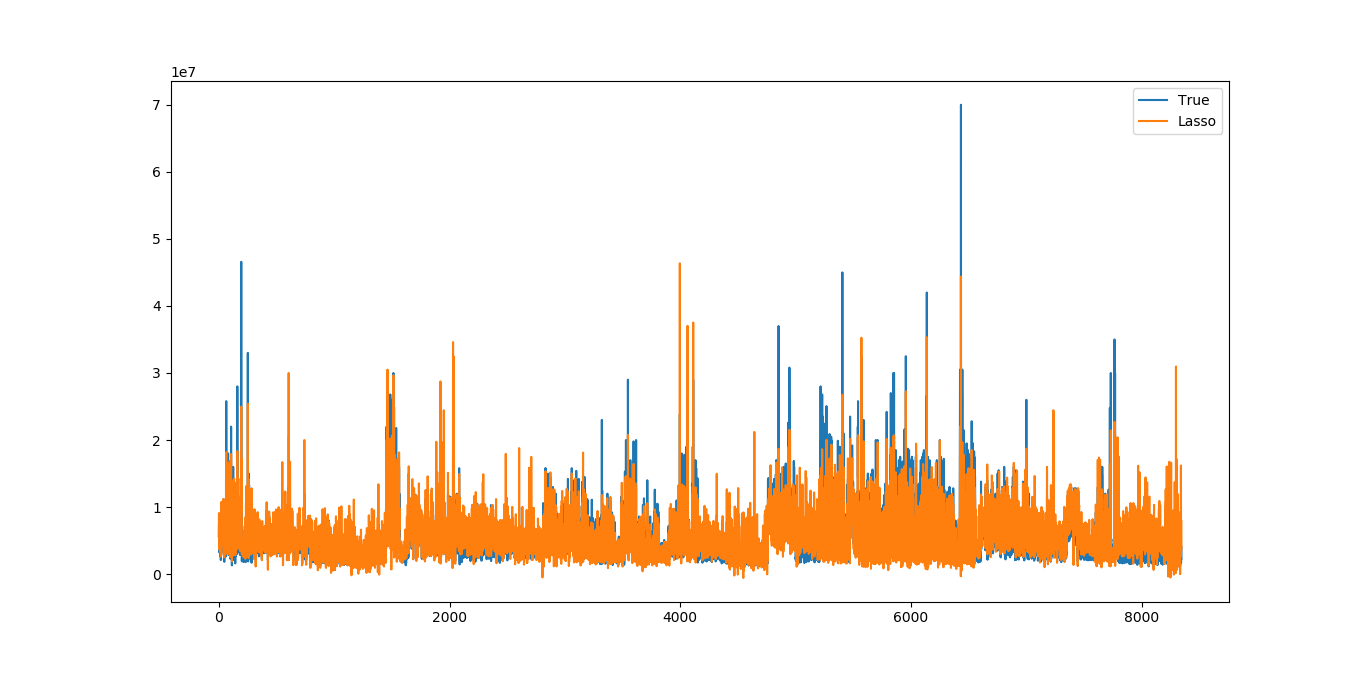
1. split.py将数据集分裂成训练数据集train\_data.csv和测试数据集test\_data.csv，各约占80%和20%
2. 训练模型并调试，主要使用sklearn库中的一些逻辑回归模型（train\_data\_sklearn.py）以及使用keras搭建神经网络（train\_data\_nn.py）。

sklearn库中使用了普通线性模型、岭回归模型、lasso模型进行预测，R^2分别为0.696975214012，0.696961885394，0.696974904752。

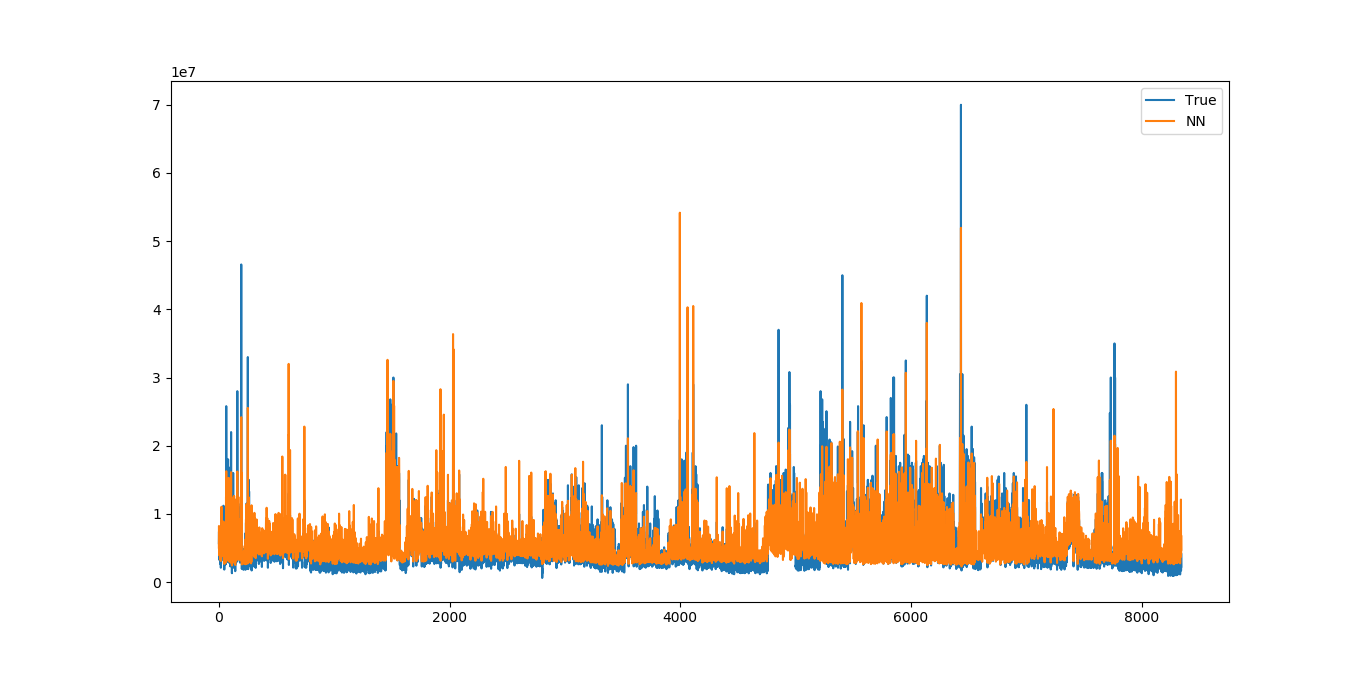
真实房价和模型回归曲线：







用keras搭建了一个2层的神经网络，激活函数使用relu，R^2为0.728964325027

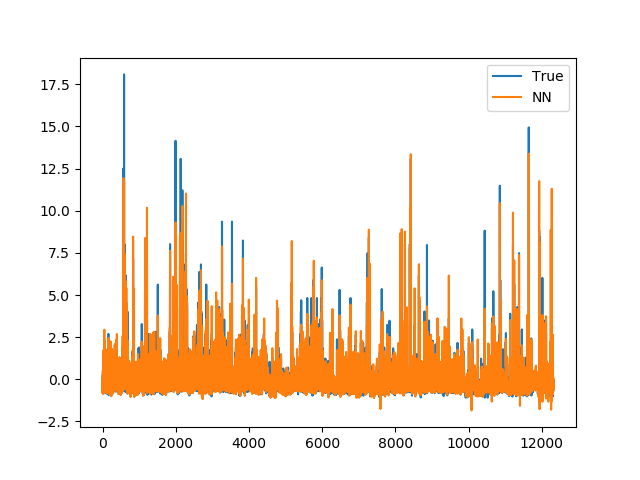


可以看出神经网络的效果更优。

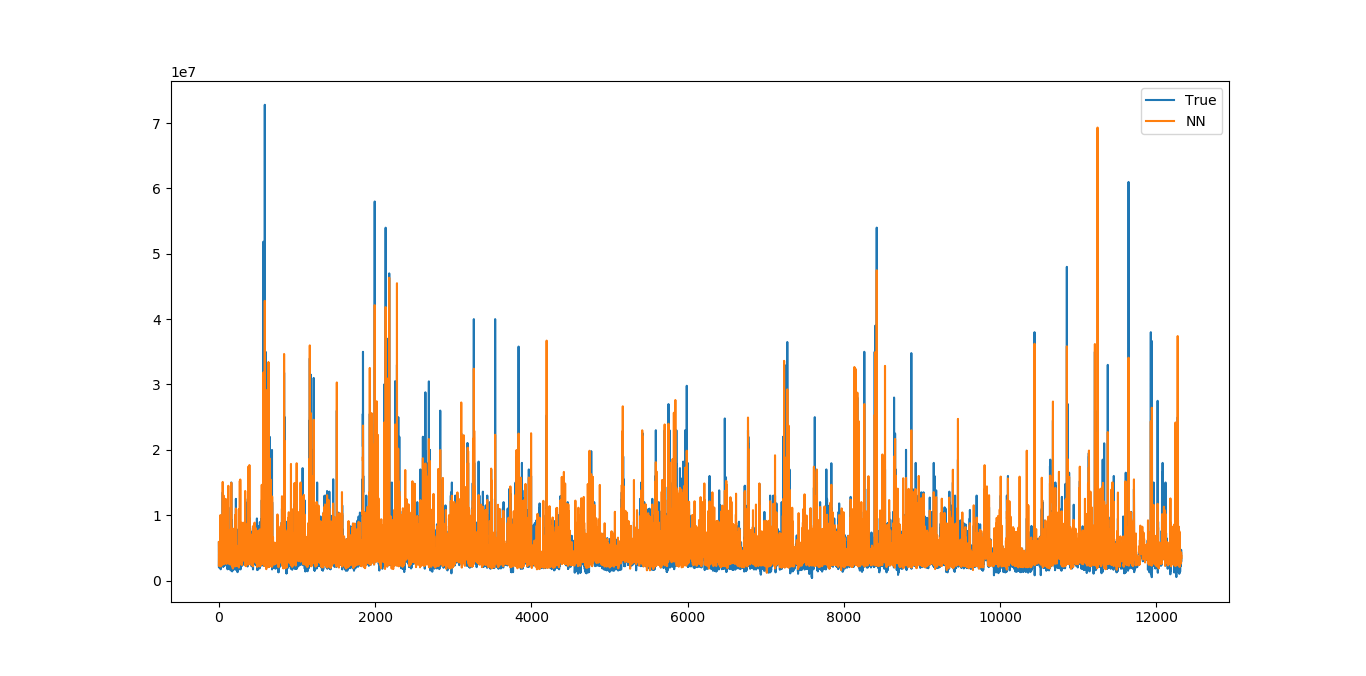
进一步改进：

由于不同变量常常使用不同的度量单位，从数值上看它们相差很大，容易使基于距离度量的模型容易受数值较大的变量影响。所以对数据进行归一化，将数据压缩到一个范围内，从而使得所有变量的单位影响一致。normal.py对data\_r.csv中的数据归一化后，保存到data\_n.csv，并重新划分训练集和测试集。

再次训练模型并预测，普通线性模型、岭回归模型、lasso模型变化不大，而神经网络模型的R^2值有了明显的提高0.748450522912



又发现神经网络模型在训练过程中的均方误差mean\_squared\_error要优于预测时的均方误差，所以可能存在过拟合的现象，所以在神经网络中添加的dropout层，在训练过程中每次更新参数时按0.05的概率随机断开输入神经元，防止过拟合。并经过进一步调参后，预测结果的R^2值能够保持在0.747以上，现module.h5为较优的模型，R^2为0.75180572067



其他文件说明：

spider.py 爬虫；

database.py初步预处理；

update.py统一格式，缺失标签补空值；

searchForCommunity.py用于将小区名称转换成坐标（坐标通过高德地图爬取POI）

coordinate.csv 不同地区，经纬度与价格对照关系，初步分析时作图使用